图书情报工作
第64卷 第12期 2020年6月

专利主体视角下专利文献与学术论文关联关系发现研究*
——以"数据挖掘"主题为例

■ 宁子晨 魏来
东北师范大学信息科学与技术学院 长春 130117

摘 要：[目的/意义]专利文献与学术论文分别表现了技术创新与科学研究的新进展，通过专利主体将二者结合进行技术主题演化分析，对进一步发现专利技术与科学研究间的关系有一定的参考意义。[方法/过程]以数据挖掘领域的学术型发明人为纽带，从专利主体－关键词耦合、IPC耦合以及IPC－关键词共现3个角度，提出关联方法并构建研究框架，分析不同时间段下主体、技术及主题多维关联关系的演化，探究数据挖掘领域内专利文献与学术论文的主体、主题关联关系。[结果/结论]学术型发明人在数据挖掘技术创新中的作用越来越重要，大多主体的技术主题是相近的，部分甚至呈现高度的统一，但是也存在少数技术与主题不直接相关，差异度较大，但不论技术与主题是否直接相关，数据挖掘相关技术发明与科学研究都已经实现了较为深入的相互渗透。

关键词：专利主体 专利文献 学术论文 关系发现
分类号：G250
DOI：10.13266/j.issn.0252－3116.2020.12.012

1 引言

新一轮科技革命和产业变革正在孕育兴起，全球科技创新呈现出新的发展态势和特征，科学技术的创新是发展的直接动力。但在我国当前科技发展实践中，科学与技术尚未呈现出良好的互动态势，主要表现在部分科学研究成果无法及时应用于技术实践，诸多技术问题往往因缺少新的科学成果而得不到有效解决[1]，一定程度上限制了科技进步与社会发展。对此，厘清科学与技术的关联关系尤为重要，陆续已有学者对科学与技术的关系及相关研究方法进行探究，旨在促进二者相互渗透、相互作用，加速科学技术的交叉融合。目前，科学研究的成果主要以学术论文的形式产出，而专利信息作为可获得的最大技术信息源，承载了技术创新的核心内容，因此科学技术关联主要体现在专利文献与学术论文的关联发现。图情领域对专利文献与学术论文间的关系陆续有所研究，早年研究多从主体间关系（发明人与作者间关系）探究科学与技术的关系[2-3]，但2010年以后对专利文献与学术论文关联的研究则重点从内容的角度出发，以更深入地探究科学与技术的关系。

近年来国外图情领域的学者对专利文献与学术论文的探索主要集中在二者的共引关系上，表现为3个方面：①通过分析专利文献中学术论文的引用情况，或学术论文中专利的引用情况，分析发现引用关系的变化[4-6]；②通过专利与学术论文的互引，探究知识在学术向技术流动过程中的促进作用[7-8]；③通过互引关系，构建引用网络，对网络进行拓扑分析，探究技术与知识在领域发展中发挥的作用[9-10]。在互引角度的研究文献较多，也取得了一定的成果。除此之外，也有学者通过技术主题演化来探究领域主题演变情况[11]，并在此基础上提出推荐系统以提高简单词检索的检全率[12]，从词的角度进行探究，提供研究参考。同时，对于专利与科技文献的关系发现不只局限于理论研究，领域学者也进行了应用研究，通过对专利文献与学术论文间多种关系的研究，发现某主题领域内的技术发展情况[13]，探究区域经济增长的情况[14]，乃至对国家、机构等的科研生产力投入产出的综合评价[15]等。

* 本文系中央高校基本科研业务专项资金项目"科学数据引用行为及其影响因素研究"（项目编号：130021822）研究成果之一。
作者简介：宁子晨（ORCID：0000－0002－3962－8490），硕士研究生，E-mail：ningzc317@nenu.edu.cn；魏来（ORCID：0000－0001－9039－1065），副教授，博士后，博士生导师。
收稿日期：2019－11－27 修回日期：2020－03－05 本文起止页码：106－117 本文责任编辑：易飞

国内关于专利文献与学术论文的研究相对国外较晚，共引关系首先成为国内研究的基础[16－17]。近几年随着本体、语义网的发展，有学者从词的角度，利用科学计量的方法，挖掘专利与学术论文的关系，如主题关联演化[18]、相似度的计算方法[19]等，提供了新的研究角度。在应用方面，除了特定领域的技术、研究发展情况的探寻，前沿热点的相关研究中也考虑到了专利文献的技术贡献[20]。

国内的专利文献缺少引文项的内容，而专利的分类方式和学术论文的分类方式也存在差异，因此不适合直接进行引文分析或分类号关联分析。本文提出从专利主体的角度出发，专利权人是专利的所有者，而专利发明人作为技术的开发人员，同时也可能是学术型研究者，基于这两类主体，可以建立创新专利与科学研究成果间的关系，通过分析专利主体－文献关键词耦合的情况以及专利主题耦合关系，进而构建专利主题与文献关键词网络，并以“数据挖掘”领域为例，分析探究“数据挖掘”主题下专利主体－技术主题－文献关键词的演化规律，希望为专利文献与学术论文的关联发现提供参考作用。

2 专利主体－关键词耦合概念界定及相关研究现状

2.1 专利主体－关键词耦合概念界定

2.1.1 专利主体的概念及界定

专利主体是与专利的形成、申请、利用等专利生命周期相关的主体，因此现实意义上，一项专利技术一般存在多个主体。比如，作为专利发明创造中的核心部分的发明人，直接参与完成发明创造，是对于发明创造作出了创造性贡献的人[21]；申请人则是指依法享有某项发明创造、向国务院专利行政部门提出专利申请的自然人、法人或其他组织[22]；而专利权人则是专利所有权的拥有者。此外，专利主体还包括专利受让人、专利代理人等多个主体。但对于一项专利，专利权人和发明人是必不可少的。当专利发明人本身也是技术的所有人时，发明人是申请人也是专利权人，但对于机构而言，专利权人一般是本机构，由相关的开发团队进行发明。

学术型发明人是特殊的专利发明人，王刚波、官建成的文章[23]中定义，“学术型发明人”来源于文献中“Academic Inventor”一词的翻译，是指在大学中既从事学术研究又从事专利活动、既具有论文作者身份又具有专利发明人身份的研究者。本文认为，具有学术创作且有学术文献产出的专利发明人，即为学术型发明人。

由于本文是通过专利主体发现专利文献与学术论文间的关联关系，因此将专利主体限定为专利权人和专利发明人，专利权人可以是机构也可以是个人，专利发明人则只能是个人。专利权人是专利的所有人，对应着专利文献；而该专利条目信息中对应的发明人，同时也可能进行着学术创作，对应着学术论文。因此通过专利主体，可以有效地建立专利文献与学术论文的关系。

2.1.2 专利主体－关键词耦合的概念及界定

在图书情报领域的研究中，学者们普遍认为关于耦合的研究最早是由美国人开斯勒提出的，如在《论“引文耦合”与“同被引”》一文中提到：美国学者开斯勒(M. M. Kessler)博士于1963年首次提出了文献耦合(Bibliographic Coupling)的概念[24]。开斯勒发现了引文耦合规律——如果A文献与B文献同时引用了C文献，则A和B之间存在耦合关系，A和B之间是存在相近关系的，这种耦合同样可以应用在专利主体－关键词的耦合上。

专利主体－关键词耦合则是专利权人与关键词之间的耦合，这里的关键词并非专利文献的主题词，而是指科技文献中的关键词。文章借助专利文献给出的专利权人与发明人，以专利发明人为中介，在中国知网及万方数据库中检索发明人发表的学术论文并记录论文的关键词，从而形成专利权人与关键词的耦合网络。通过专利发明人建立专利主体(专利权人)与关键词的耦合关系，以找到与该专利相关性较高的科技文献，并提供对应的关键词以用作相关分析。

2.2 可行性分析

2.2.1 国内专利少有引文，引文关联研究存在盲区

在基于专利文献与学术论文的关联发现研究中，虽然国内外的研究手段主要集中在引文网络方面，但是中国的专利数据库基本没有引用专利，仅有最近几年专利审查员添加的极少数引用专利[25]。同时，专利文献与科技文献的体例范式不同，内容的侧重点也有所不同，这使得学者在学术创作中很少引用专利文献，在专利引用的相关研究中会形成一定的盲区。

2.2.2 主题作为知识单元，能较好反映主体的学术背景

在科学知识表达中，主题作为中观层次的知识单元，它是从内容的角度代表了该文献作者的特定研究领域或学科背景。在科学与技术的知识网络中，虽然

主题单元是表征知识内容的隐性科学分子,但是与其他知识单元关联密切,有着影响科学发展历程与趋向的作用[26]。虽然主题可以从分类的角度来揭示,但是专利文献与学术论文的分类体例相差甚远,无法全面合理地建立映射关联,因此仍需要从词的角度出发,在内容上将相似的节点关联,进而发现专利文献与学术论文的关系。

2.2.3 专利主体是专利文献与学术论文关系构建的基础

专利的分类是按照其应用领域进行分类的,而科技文献是按照中图分类法进行分类的,二者有很大的差别,因此不能直接通过主题词进行映射关联。主体作为技术、知识的承载者,在促进科学技术间知识流动上发挥着重要作用。在专利发明创造中,参与的发明人可能是学术型发明人,通过跟踪该类发明人的学术论文,与专利文献建立联系,能较好地反映技术发展中知识的流动情况。因此基于专利主体,从技术－主题角度探究专利文献与学术论文的关联关系及其演化规律,能较为容易地从专利主体的学术背景和技术创新的角度来实现研究目标。

因此文章基于专利主体(以学术型发明人为主),以专利 IPC 与学术论文的关键词为研究对象,进而探究专利主体与文献主题间、专利主体与技术间以及技术与文献主题间的关系,发现技术主题演化的规律。数据可获得,能够较好地实现技术领域内专利文献与学术论文间关联关系的发现。

3 关联方法及框架构建

专利主体既参与了专利的发明创新,又进行了学术创作,因此其对应的专利文献与学术论文之间主题较为相近。以专利文献与学术论文为基础,以专利主体为中介,从专利主体－关键词耦合、技术演化以及技术主题关联,来探究数据挖掘领域专利文献与学术论文间的关联关系。其整体研究框架如图 1 所示,一项专利技术对应一个或多个专利权人,进而对应多个发明人,而部分发明人同时进行了学术创作,同时期公开发表了学术论文,因此专利文献与学术论文间可基于专利主体建立一定的联系,即"专利文献－专利权人－专利发明人－学术论文"的关系。本研究中,首先对既定技术主题进行专利文献的检索,记录详细数据;之后通过专利信息,形成相应的技术共现网络;同时,通过专利主体检索相关的学术论文,并记录论文关键词,形成专利主体－关键词耦合网络;最后,通过专利主体,构建起专利技术主题与文献关键词间的关系网络,进而发现专利文献与学术论文间的关系。

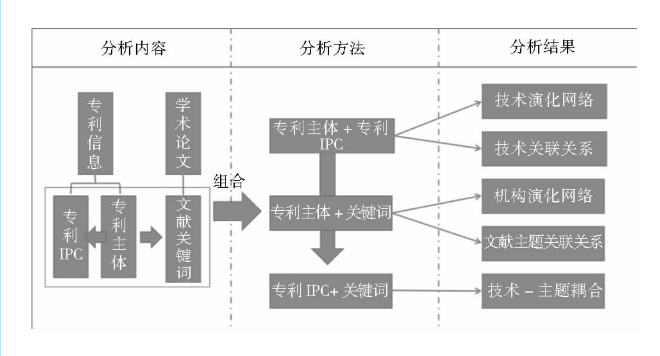


图 1 整体研究框架

3.1 专利主体－关键词耦合网络

专利主体与文献关键词的关联,实则是以学术型发明人为纽带,指向发明人对应的专利文献以及发表的学术论文。笔者基于中国知网和万方数据库检索了专利发明人在专利申请的前 2 年内发表的学术论文,记录该项专利的专利权人以及对应发明人发表学术论文的关键词,进而构建专利主体与文献关键词的耦合网络,如图 2 所示。由于一般来说发明专利自申请日起 18 个月后公开,因此本文选择了发明人在专利申请日前 2 年内发表的学术论文,并记录论文关键词,这些关键词一定程度上可以反映发明人的技术背景或者在专利发明期间的主要技术方向。

一条专利记录对应专利权人 1 和专利权人 2,专利权人 1 对应发明人 11、发明人 12 以及发明人 13,专利权人 2 对应发明人 21、发明人 22,不同的发明人又对应不同的文献关键词 a－h,最终形成了一项专利的专利权人与关键词间的耦合关系。

3.2 技术共现网络

技术的耦合情况,主要从专利 IPC 分类号的耦合来表示。专利的分类是按照其应用领域进行分类的,因此只包括 8 个大部,采用等级的方式,对部－大类－小类－大组－小组进行逐级分类,如图 3 所示(IPC 号具体含义可见国家知识产权局－中国专利公布公告:http://epub.sipo.gov.cn/ipc.jsp)。

一个专利可能有多个 IPC 分类号,同时涉及多个技术领域,对专利的 IPC 分类号进行共现处理,可以反映特定主题下的核心技术领域、边缘技术领域以及各技术领域之间的关系。之后利用 PageRank 对节点进行统计处理,通过对边的计算来反映网络中节点代表的技术领域的重要程度。

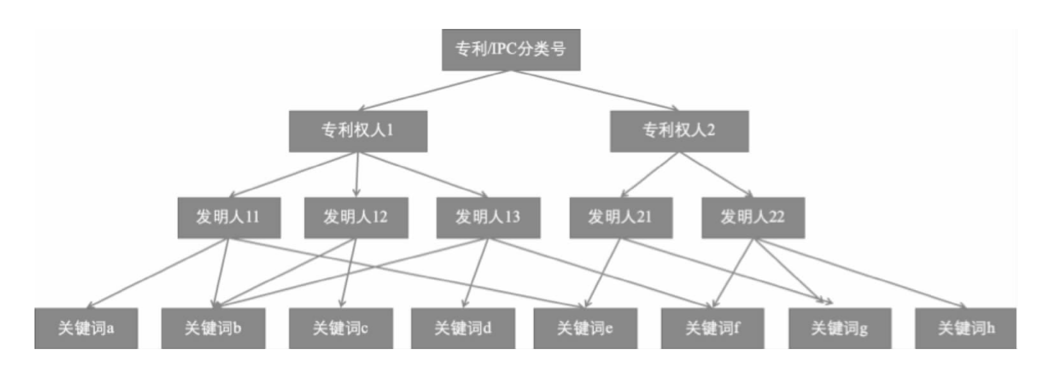


图 2　专利主体与关键词耦合关系

Gephi 对 PageRank 算法做了修正，减弱了“自私互指关系节点”以及“无出入度节点”等对网络中关系的影响[27]，可以在呈现专利 IPC 耦合关系的基础上，反映专利 IPC 的重要程度，进而从专利的层面分析技术的关联情况。

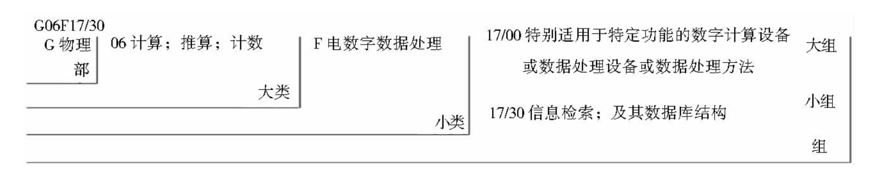


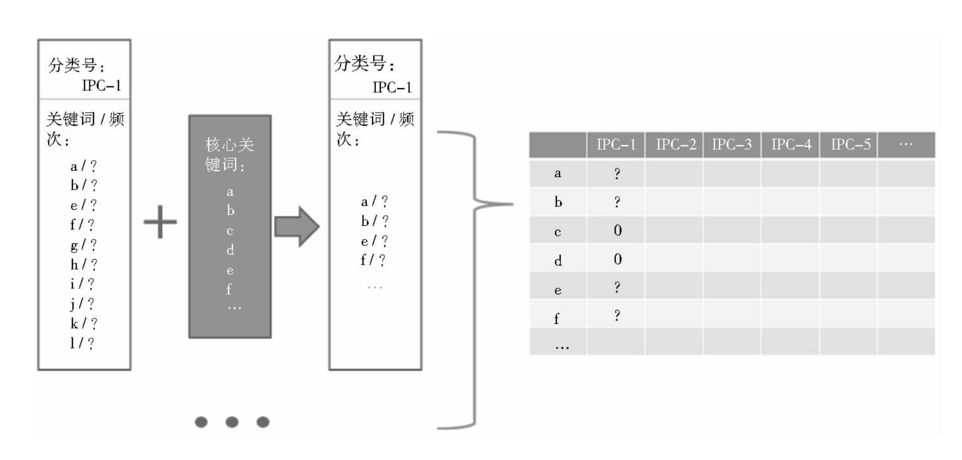
图 3　专利 IPC 分类号及示例

3.3　专利技术－文献关键词关系网络

技术主要是通过专利 IPC 来表现，关键词则是学术论文的关键词，技术－主题网络实则是 IPC 分类号－关键词共现的结果。由于参与一项专利创新的发明人人数不定，则一条专利记录对应的技术背景涉及多个方面。同时这些发明人在对应时间内的发文情况也不确定，这导致了一个专利对应的关键词的离散程度可能很高，因此需要首先确定关键词的范围，进而构建技术－主题网络。

在关键词范围的选择上，利用 Gephi 的 K－核心对专利主体－关键词耦合网络的数据进行过滤，通过显示度来判断是核心关键词还是边缘关键词。核心关键词可以有效地涵盖研究领域中知识点的整体分布，同样地，它也可以反映出某一项专利的主要技术方向或学科所属，因此核心关键词的选择是必不可少的。边缘关键词虽然连接度低，但内容多样，一定程度上可以反映弱相关的主题与技术之间的隐性关系，因此也需要考虑这类词的变化情况。

关键词确定之后，基于专利主体构建专利 IPC 与关键词的共现网络。将同一时间段下同一专利 IPC 与核心、边缘关键词分别聚合，重复的关键词进行频次累加处理，如图 4 所示。最终可以得到所有核心边缘关键词与 IPC 分类号的对照关系，即形成 IPC 分类号－核心关键词矩阵，将其可视化结果进行模块化处理，模块间耦合度较低，模块内聚合度较高，进而可以对技术与主题间的关系进行探索。



	IPC-1	IPC-2	IPC-3	IPC-4	IPC-5	…
a	?					
b	?					
c	0					
d	0					
e	?					
f	?					
…						

图 4　IPC 分类号－关键词耦合示例

4 基于主体的专利文献与学术论文关联发现实证分析

4.1 数据的选取

大数据因为近年来互联网和信息行业的发展而引发人们的关注，有效地对数据进行收集、分析、保存以及共享，可以为科研提供有效的帮助，为用户解决切实的问题[28]。本文以“数据挖掘”为主题，在中国专利局“专利检索与分析”数据库中进行检索，利用布拉德福定律确定核心专利权人（机构）22 个以及相关专利数据 248 条。爬取详细的专利数据，包括专利题名、申请号、申请人、公开号、公开日、IPC 分类号、申请人以及发明人等数据。

为了方便分析科学技术的演变情况，将 248 条专利数据按 2004－2008 年、2009－2013 年、2014－2018 年 3 个时间段分为 3 组，并分别编号为“1001、1002…2001、2002…3001、3002、3003…3150”。其中，首位数字代表所属时间段，如“1xxx”代表在 2004－2008 年内的专利数据。

之后对这 248 条专利数据中的专利发明人发表过的学术论文进行检索：在中国知网以及万方数据库中，以专利发明人的姓名为“作者”、专利权人的名称为“单位”、对应专利申请时间的前 2 年为“发表时间”进行限定检索，下载全部的题录信息并进行合并整理去重，同样进行编号“1001、1002…2001、2002…3001、3002、3003…3150”。

通过编号关联整合专利信息和学术论文的数据，将数据导入 Gephi0.9.2 中，分 3 个时间段进行个别和整体分析。

4.2 结果

4.2.1 IPC 共现结果

利用 BibExcel 对数据进行处理，同时利用 Gephi 的 PageRank 进行成图，得到 2004－2008 年、2009－2013 年、2014－2018 年“数据挖掘”相关专利分类号的耦合情况，如图 5－图 7 所示，各耦合网络的相关参数如表 1 所示。

由于本文对 IPC 分类号作耦合化处理，因此网络中的“孤立点”被剔除，仅留下了存在关系的节点。从网络参数上来看：①3 个时间段的节点数、边数以及平均加权度都是增加的，这符合技术的发展情况，多技术相融合，共同进行创新。②连接组件是指子图中的各个节点间通过边相连，但是子图间是不存在关系的。3 个时间段的连接组件参数很小，但略有增加，节点/连接组件的值也是增长趋势，可见耦合关系网络是增长的，子图也是扩张容纳更多的技术节点。③3 个时间段下的图密度很小，且随着时间不断减小，说明耦合网络关系的增长与节点的增长是不匹配的，虽然节点和边都增加了，但实际构成的耦合网络是越来越稀疏的。此外，平均路径长度的增加也说明了网络是愈加稀疏的。

2004－2008 年间的 IPC 通过中心节点 G06F17/30 建立关系，形成一个子群，节点间的耦合度基本一致。在 2009－2013 年间，节点间的关系变复杂，在以 G06F17/30 为中心节点的子群之外，又出现了 2 个新的子群。在 IPC 的耦合度方面，G06F17/30 与 H04L29/06（08）、G06N3/12 以及 G06F17/50（27）有着较强的耦合度。而在 2014 年之后，专利 IPC 的耦合关系更加复杂，中心节点变为 G06F17/30 以及 G06Q50/06，子群的个数以及子群内部的节点数相对 2009－2013 年间有所增加。这一阶段中，G06Q50/06 与 G06Q10/06 的耦合度最高，其次是 G06Q50/06 与 G06F17/30、G06Q50/06 与 G06Q10/04，节点耦合度高，说明一项专利技术同时涉及这几个节点所代表的领域，应用范围较广。专利 IPC 的增加以及网络愈加复杂，一方面说明“数据挖掘”越来越多地被应用在多个领域，另一方面也体现了多领域的技术、知识被运用在数据挖掘中，是学科融合的体现。

表 1 专利 IPC 耦合网络参数

度量指标	2004－2008 年	2009－2013 年	2014－2018 年
节点	4	15	44
边	4	13	61
平均加权度	2	2.4	5
网络直径（D、R、APL）	2、1、1.333	3、1、1.857	5、1、2.482
图密度	0.667	0.124	0.064
连接组件（孤立点）	1	3（7）	6（14）
平均聚类系数	0.778	0.342	0.808
特征向量中心度	0.000 016 43	0.000 688 44	0.002 982 25

注：表中 D、R、APL 分别代表直径、半径、平均路径长度

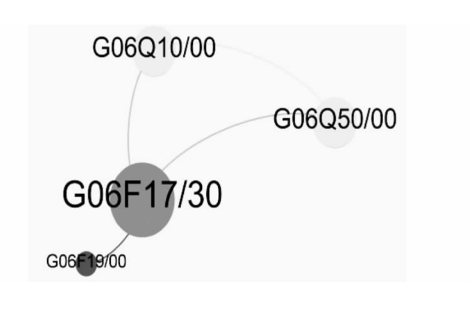


图 5 2004－2008 年专利 IPC 耦合网络

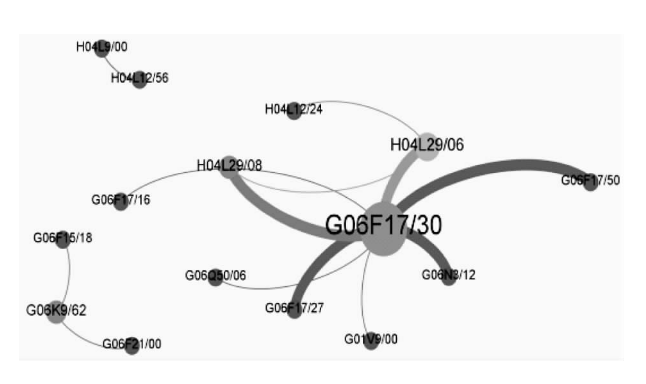


图 6　2009 - 2013 年专利 IPC 耦合网络

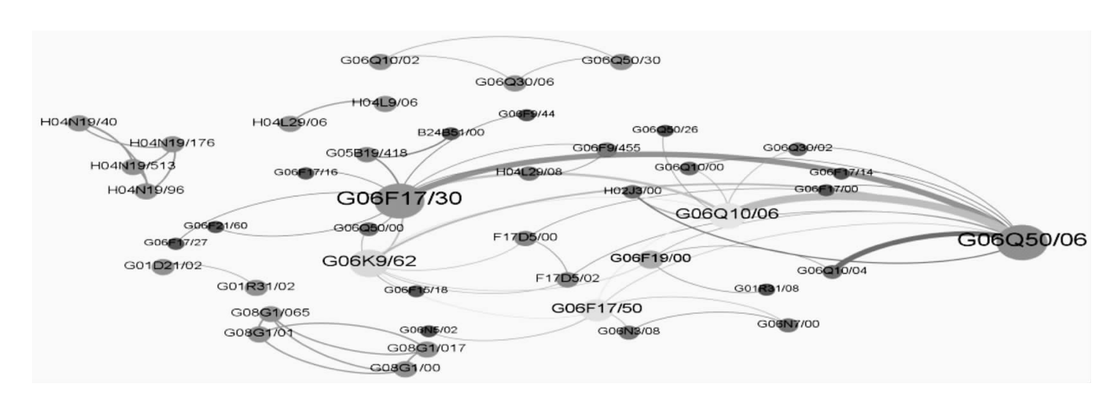


图 7　2014 - 2018 年专利 IPC 耦合网络

4.2.2　专利主体 - 关键词耦合结果

利用 Gephi 呈现 2004 - 2008 年、2009 - 2013 年、2014 - 2018 年“数据挖掘”相关专利主体 - 关键词的耦合情况，如图 8 - 图 10 所示，各耦合网络的相关参数如表 2 所示：

表 2　专利主体 - 关键词耦合网络参数

度量指标	2004 - 2008 年	2009 - 2013 年	2014 - 2018 年
节点	94	714	5 998
边	103	904	14 413
平均加权度	2.191	2.532	4.806
网络直径(D、R、APL)	5、3、3.044	8、4、4.162	7、4、3.544
图密度	0.024	0.004	0.001
模块个数	3	11	18
模块度	0.564	0.814	0.617

注：表中 D、R、APL 分别代表直径、半径、平均路径长度

从网络参数上来看：①3 个时间段的节点数、边数的增长迅猛，平均加权度也是增加的，这表明了随着时间发展，专利主体增多，专利涉及的研究主题也增多且趋向融合。②网络模块解析度的值为默认值 1.0，得到模块个数由最初的 3 个增长到 18 个，节点标签的大小表示该节点在网络中的重要程度，子群内部相关性更高。③3 个时间段的图密度很小，且随着时间不断减小，说明虽然网络的节点和边都增加了，但实际构成的耦合网络是越来越稀疏的。

2004 - 2008 年间的 3 个核心专利主体是清华大学、上海交通大学以及浙江大学，3 个主体间的共有关键词较少，仅有“数据融合”“特征选择”与“数据挖掘”。在 2009 - 2013 年间，模块个数增加到 11 个，核心节点增加了国家电网、重庆大学、南京邮电大学等，但是清华大学在这一阶段的表现不显著。模块间的关系也变得复杂，相对于图 8，图 9 的主体间的关系增多，独立性降低。共有的关键词增多，包括“网格计算”“支持向量机”“分布式计算”“监控系统”“IEC61850”“综合应用服务器”“谐波治理”“智能变电站”“认知无线电”等。在 2014 年之后，专利主体 - 关键词耦合网络明显复杂化，国家电网公司占据了网络最核心的位置。其次，各所邮电大学也逐渐崭露头角，成为核心节点。共有关键词的频次和数量都所有增长，高频共有关键词有“大数据”“神经网络”“多目标优化”“支持向量机”“关联规则”“储能系统”“层次分析法”等。网络趋于复杂，专利主体的增长变化体现了领域内机构的变化，而关键词节点的增长则体现了技术领域的发展，耦合度高的关键词则在网络中的重要程度更高，在促进技术发展中发挥着更为重要的作用。

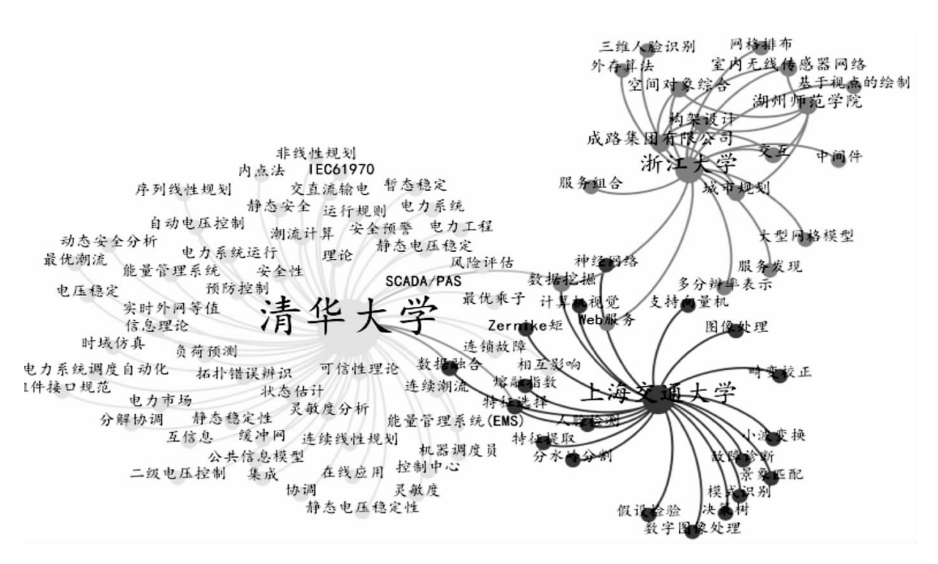


图 8 2004 –2008 年专利主体 – 关键词耦合网络



图 9 2009 –2013 年专利主体 – 关键词耦合网络

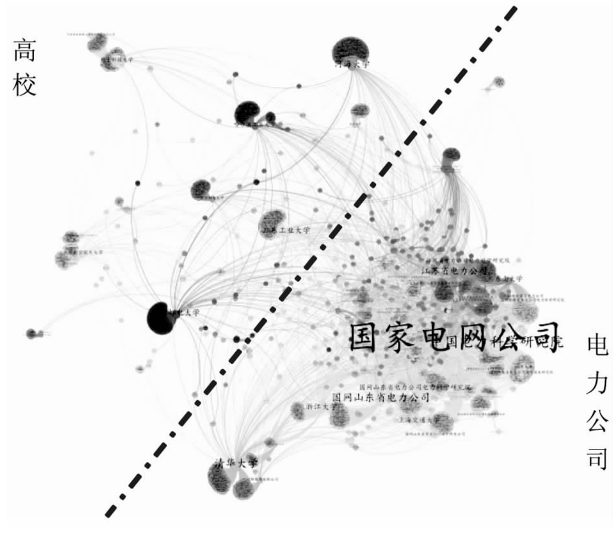


图 10 2014 –2018 年专利主体 – 关键词耦合网络

4.2.3 IPC –关键词共现结果

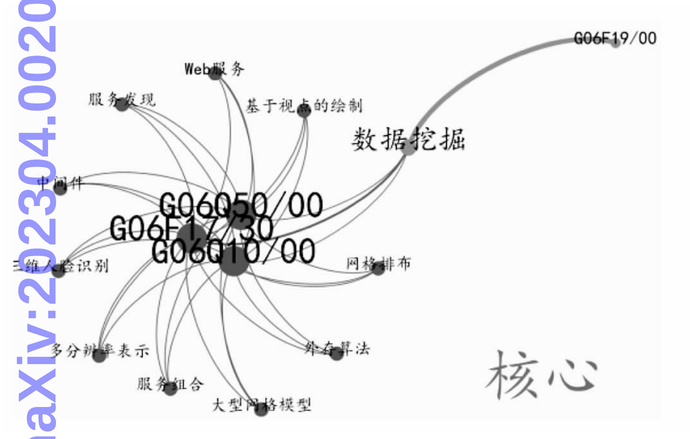
利用 Gephi 呈现 IPC – 核心关键词网络和 IPC – 边缘关键词网络，如图 11 – 图 13 所示，节点的大小代表该点在网络中的重要程度，而节点的灰度则是模块化的区分。

IPC – 核心关键词共现结果为：在 2004 –2008 年间主要形成 2 个模块，2009 –2013 年间形成 3 个模块，2014 –2018 年间形成 7 个模块；IPC – 边缘关键词共现结果为：在 2004 –2008 年间主要形成 2 个模块，2009 –2013 年间形成 5 个模块，2014 –2018 年间形成 13 个模块。边缘共现网络相较核心共现网络更加分散，形成的子群间缺少联系，这在模块个数上也得到了体现。IPC 和关键词的分布情况如下：

（1）IPC – 核心关键词共现网络 2004 –2008 年间，

IPC 主要为 G06Q50/00、G06Q10/00 和 G06F17/30，对应的技术领域至小类为物理部 – 计算推算计数大类 – 电数字数据处理小类和数据处理系统或方法小类；关键词主要是网络开发，服务发现相关。2009 – 2013 年间，IPC 主要为 G06F17/30、H04L29/06 以及 G06F21/55 等，对应的技术领域至小类为物理部 – 计算推算计数大类 – 电数字数据处理小类、电学部 – 电通信技术大类 – 数字信息的传输小类；关键词则主要涉及网络安全、服务器承压能力控制相关。2014 – 2018 年间，IPC 主要为 G06F17/30、G06Q50/06、G06Q10/06 以及 G06K9/62、H04N19 等，新增了电学部 – 电通信技术大类 – 图像通信小类以及物理部 – 计算推算计数大类 – 数据识别小类等技术；关键词多偏向应用，如交通、移动通讯、电力、能源、信息检索等方面。

(2) IPC – 边缘关键词共现网络 2004 – 2008 年间，IPC 主要为 G06F19/00 和 G06F17/30，涉及的关键词主要是计算机视觉、应用方向(电力、能源)、服务器承压能力控制、服务发现等方面的术语。2009 – 2013 年间，IPC 主要为 G06F19/00、G06F17/30、G05B13/04 以及 G05B19/418 等，对应的技术领域新增了为物理部 – 控制调节大类 – 一般的控制或调节系统小类；关键词则主要是测试、网络安全等必要技术相关术语以及人工智能、Web 服务、具体应用相关的术语。2014 – 2018 年间，IPC 主要为 G06F17/30、G06Q50/06、G05B23/02 以及 G01R31/12 等，G01R 为物理部 – 测量测试大类 – 测量电磁变量小类；关键词涉及了机器学习、深度学习、计算机视觉等技术术语和网络安全防护、网站维护等必备的基础术语，以及各领域的应用相关技术词汇(如金融、电气、智慧城市、交通等)。



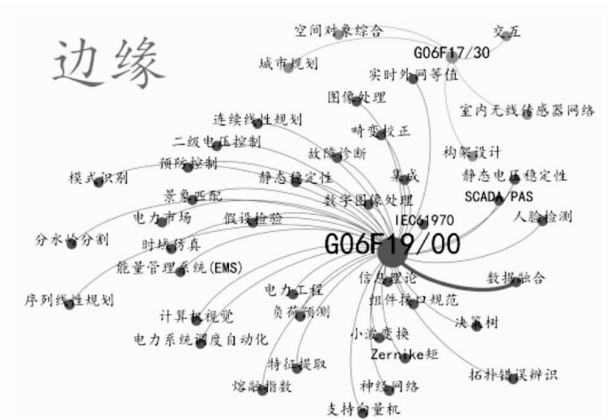
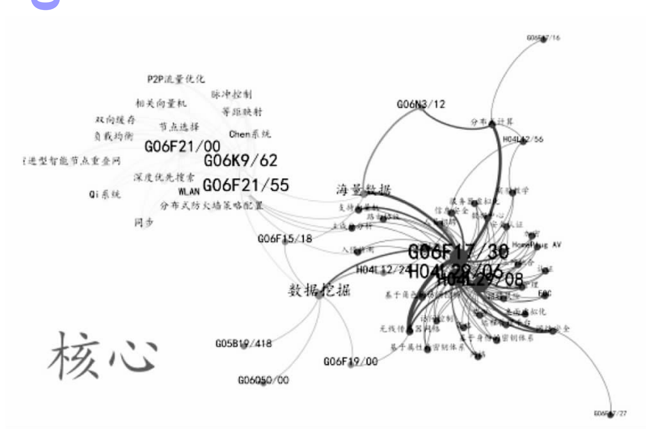


图 11　2004 – 2008 年 IPC 分类号 – 关键词共现网络



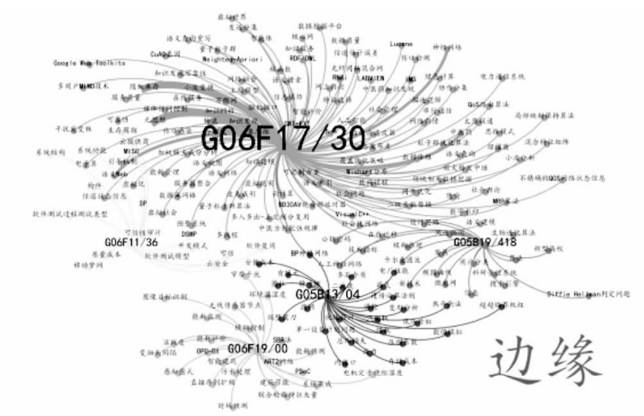


图 12　2009 – 2013 年 IPC 分类号 – 关键词共现网络

4.3　基于主体的专利文献与学术论文关联关系分析

针对形成的网络图以及呈现的结果，对其关联关系做进一步的分析。

4.3.1　技术演化分析

内容方面，在专利分类号的种类上，3 个时间段分别为 4 种、15 种、44 种，各属不同的大部。2004 – 2008

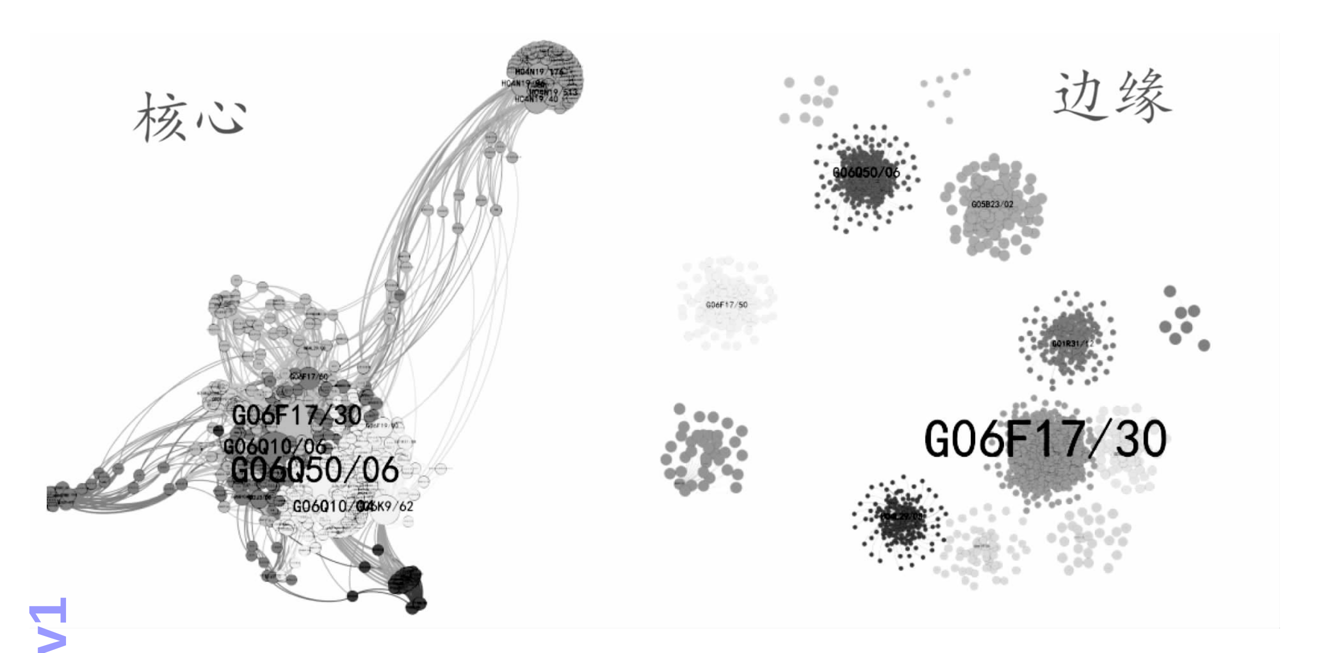


图 13 2014 – 2018 年 IPC 分类号 – 关键词共现网络

年间主要集中分布在物理大部下的“计算;推算;计数”类发明,2009 – 2013 年中增加了物理大类的“测量;测试”类发明与电学大类的“电通信技术”相关专利技术,在第 2014 – 2018 年间,更是拓展到物理大部的“信号装置”“控制;调控”、电学大部的“发电、变电或配电”、作业运输大部的“磨削;抛光”以及机械工程大部的“照明;加热;武器;爆破 – 气体或液体的贮存或分配”相关的专利技术创新。可见,越来越多的应用领域开始重视数据挖掘的利用与创新。

在核心节点上,在 2014 年以前的 2 个时间段里,G06F17/30 都处于核心位置,是最重要的节点。而在 2014 年之后,核心节点在 G06F17/30 的基础上,又增加了 G06Q50/06。虽然技术上均属于 G06 物理部下的“计算;推算;计数”大类,但是前者主要是电数字数据处理中的“信息检索;及其数据库结构”相关技术创新,后者是专门适用于行政、商业、金融、管理、监督或预测目的的数据处理系统或方法中的“电力、天然气或水供应”相关技术。G06F17/30 仍是计算机网络技术、数学算法这类“数据挖掘”的主流技术,而 G06Q50/06 作为核心的出现,表示“数据挖掘”相关技术创新不再仅集中于计算机领域,在电力水利检测方面也涉猎颇广。

技术耦合演化方面,在 2009 – 2013 年间,G06F17/30 与 H04L29/06(08)、G06N3/12 以及 G06F17/50(27)有着较强的耦合度,即“信息检索及其数据库结构”与“通信控制通信处理”常出现在一个专利中,“基于遗传模型的的计算机系统”与“计算机辅助设计(自动分析的自然语言处理技术)”常出现在同一专利发明中。在 2014 – 2018 年间,G06Q50/06 与 G06Q10/06 以及 G06F17/30 形成较为稳定的耦合关系,即“信息检索及其数据库结构”“行政管理相关的数据处理系统或方法”以及“适用于电力、天然气或水供应经营部门的系统或方法”间的关联度较高。2009 – 2013 年的技术耦合,主要以数据挖掘相关的技术为主,而 2014 – 2018 年更重视技术与领域应用的结合,这一过程中,技术发明创新的重心由相关技术的优化改进发展到特定领域内的特定应用,取得了一定的成效。

4.3.2 基于主体的一维关系发现

由图 8 – 图 10 的演变来看,在 2004 – 2008 年间,核心主体为清华大学、上海交通大学以及浙江大学;2009 – 2013 年间的核心主体是上海交通大学、重庆大学、南京邮电大学,此阶段电力公司集合作为重要节点出现;2014 – 2018 年间,国家电网核心成为最核心的主体,占据图 7 将近 1/2 的位置,与其强联系的模块则以电力公司为主。

由于专利主体 – 关键词耦合关注的是学术型创新人才,即具有双重身份的技术人员,因此会忽略掉一部分只做技术创新的发明人,所以这种核心主体的演变并不能直接说明数据挖掘技术最先是从高校发展的。另一方面,经过对实证数据的计算,学术型发明人占比从 2004 – 2008 年的 30% 变为 62.26% 再到 2014 – 2018 年的 77.33%,在一定程度上可以看出学术型发明人在技术发明中的占比越来越大。

在最初(2004 – 2008 年间)数据挖掘相关专利申

请的主体中，同时注重学术创新的主要是高校的创新者，这类发明人占该阶段全部主体的 30%；在技术初步发展阶段(2009 – 2013 年间)，企业(如电力公司等)开始重视专利主体的学术背景，作为一个子群出现在耦合网络中，该阶段发表过学术文献的发明人占比为 62.26%；而在数据挖掘相关技术迅速发展(2014 – 2018 年间)的进程中，拥有学术型创新人才的企业已经可以与高校分庭抗礼，而该阶段有论文发表的发明人占到 77.33%。在数据挖掘相关技术的整个发展创新中，学术型创新人才由不足 1/3 增长到接近 4/5，可见数据挖掘相关的科学研究与技术发明的关联是不断增强的，具备科研能力的创新型人才对专利活动产生了重要的影响。同时，以电力行业为主的企业从无到有，体现了产业与学术界之间逐渐形成了较高的知识能力互补。

4.3.3 专利主体 – 关键词的二维关系发现

在关键词演化过程中，2004 – 2008 年间主要是计算机网络、计算机视觉以及服务器相关的关键词；2009 – 2013 年间，在前一阶段的基础上增加了服务器优化、网络安全以及分析类关键词；2014 – 2018 年间，核心关键词的种类大大增多，主要集中在数据分类、算法相关、人工智能(机器学习、深度学习、神经网络)、电力电场等相关关键词(见图 8 – 图 10)。同类词的表达多样，内容上更加细致。

在主体 – 关键词耦合中，由于给定的关键词可能相似或相近，形成了共关键词的现象。不同主体间的共有关键词有所偏向：高校间的共现关键词偏向于理论研究，而企业与高校间的共有关键词除了相关技术类名词之外，更倾向于应用领域；企业间的共有关键词则主要以应用类为主。不同时间段呈现出的共关键词也各有特点：2004 – 2008 年间，共关键词为“数据融合”“特征选择”“城市规划”“空间对象综合”；2009 – 2013 年间，共关键词为“支持向量机”“分布式计算”“一体化监控系统”“IEC61850”“综合应用服务器”“谐波治理”“智能变电站”“认知无线电”等；2014 – 2018 年间，共关键词为“大数据”“神经网络”“多目标优化”“支持向量机”“关联规则”“储能系统”“层次分析法”等。

随着时间的发展，主题与主体的关联越来越多，专利主体重视的主题相似度也在增加，说明这些学术主题可能与该阶段相关发明创造的核心技术有关。同时，除了与企业应用相关的关键词，其他共有关键词大多是该时间段下较为热门的技术主题，这些技术主题很快被多个主体关注并应用，可见数据挖掘技术积极吸纳新的科学、技术知识，迅速发展。在 2014 – 2018 年间，一些其他学科(如运筹学等)的相关主题也成为颇为高频的共关键词，说明这类关键词在该时间段内发挥了重要的作用，一定程度上促进了数据挖掘相关的发明创造。可见在技术创新的过程中，会出现与之不直接相关的学科主题，直接或间接地影响技术问题的解决。

4.3.4 基于主体的技术 – 关键词网络关系演化分析

不同阶段的核心主题有着明显的区别，如图 11 – 图 13 所示，在 2004 – 2008 年间，涉及的技术术语并不是数据挖掘的主要技术，但却是在技术实现过程中可能使用的相同或相近的开发语言，为数据挖掘技术创新提供支持；服务发现相关主题，可能是为了提高服务，吸引受众，从而提出数据挖掘的需求。在 2009 – 2013 年间，主题主要是网络安全和服务器承压能力控制，此时数据挖掘技术基本成型，需要深入考虑优化方向的问题。2013 – 2018 年间，主题多以应用领域为主，其中电网相关技术术语最多，同时也涉及了图像识别这类偏技术理论方向的专业术语。

不同阶段的边缘主题相似度较高，都涉及了网络安全、服务器维护这类必备的基础性技术术语以及需求类、应用领域的相关术语。不同之处则在于在 2004 – 2008 年间，数据挖掘刚刚起步，边缘主题提出了机器学习、深度学习等人工智能相关的算法，而这类主题在 2014 年以后成为核心主题。可见一项技术的创新改进，需要有一定的积累与准备，2014 – 2018 年的边缘主题，在几年后同样可能成为核心主题，成为数据挖掘的核心技术。

由于上述特点，选择基于专利主体的 IPC – 核心主题的共现网络，分析发现技术主题关联网络的演化情况，更能反映专利核心技术与学科主题的关联关系。

(1) 2004 – 2008 年间，数据挖掘涉及的技术领域主要为电数字数据处理技术、行政管理金融的数据处理系统或方法、信息检索及其数据库结构技术，这些技术与科学发展中的网络开发、服务发现相关联，这些科学主题实际是在数据挖掘技术形成过程中提出的一些需求主题，与行政、管理、金融、商业等领域相匹配。

(2) 2009 – 2013 年间，G06F17/30、H04L29/06 与网络安全类主题相关联，涉及的技术领域为信息检索及其数据库结构技术以及通信控制与处理技术。在网络通讯中，数据对于开发者来说是透明的，因此重视网络安全、防止信息泄露是至关重要的。G06F21/00、

G06K9/62 以及 G06F21/55 与服务器承压能力控制主题相关联,涉及的技术领域为保护计算机及其部件、程序或数据的安全装置以及数据识别技术,服务器承压能力包括最大并发用户数、吞吐量、容灾恢复等。虽然这些科学主题与技术有一定的差异,但都是为了对数据挖掘做进一步的优化,提出优化的方向。

(3)2014 – 2018 年间,H04N19/xx 与图像识别主题相关联,涉及的技术为图像通信技术,其发明人拥有图像识别的学科背景,使得技术发明与科学发现良好地融合渗透,是数据挖掘的重要组成部分。G06F、G06N、G06Q、G01R 以及 H04L 与交通、教育、移动通讯等应用领域主题相关联,G06Q、H02J、G01R、G05B 以及 G06F 与电网相关应用领域的主题关联,G06Q、G06F、G06K、G06R 以及 G08B 与能源应用相关主题存在关联,这些技术主题共现集中在各个应用领域,电力能源行业应用尤为突出。科学知识与技术创新共同实践于应用领域,促进了知识在科学与技术中的传递。F17D5/00 与光纤、光力效应、设备等物理材料相关主题关联,该技术为保护装置或观测装置,二者都是聚焦于物理硬件的保护改良,这一技术与主题具有极高的相似性。B24B51/00 与 G05B19/418、G08G1/xx 与信息检索(自然语言处理、关联映射等)及工业生产相关主题存在关系,相关技术为用于磨削或抛光的机床装置或工艺、道路车辆的交通控制系统。这一技术主题间的相似度不高,但拥有这类科学背景的学术型发明人进行了作业运输领域的相关创新,可见这类科学知识对作业运输领域的数据挖掘创新起着重要作用。

在数据挖掘的 15 年间,G06F17/30(物理部 – 计算推算计数大类 – 电数字数据处理小类 – 信息检索及其数据库结构大组)一直都是其核心创新领域,相关学术型发明人的学科背景随着数据挖掘核心主题的变化而变化,有些主题是直接相关的,有些主题不直接相关,但是都解决了问题,推动了技术创新的发展,促进了技术在领域内的应用。在 2004 – 2008 年、2009 – 2013 年、2014 – 2018 年 3 个时间段内,电力系统、算法、数据库技术、神经网络、无线传感、网络安全等关键词是不变的,且随着时间的发展不断更新补充,可见这类学术背景的发明人在该类专利发明中的作用是重要的。

5 结论

本文通过对专利主体、专利技术、学术论文关键词的多维共现分析,讨论了专利主体关系网络、专利主体 – 关键词关系网络以及技术 – 关键词关系网络的演化情况,探究了基于专利主体的专利文献与学术论文间主体、主题的关联关系。在数据挖掘技术发展中,学术型发明人发挥着越来越重要的作用,有学术成果的发明人将其特有的学科知识运用在技术创新中,一定程度上促进了技术创新。而由 4.3.4 的技术 – 关键词演变关系中可知,在技术主题网络演化的过程中,大部分学术型发明人的科学发现与其技术创新是相近的,部分技术主题间呈现高度的统一;同时也存在少数技术与主题不直接相关,差异度较大,但在一定程度上对数据挖掘技术的发展起到了促进的作用。随着时间的推移,数据挖掘相关科学发现展现的学科领域越来越广,而相应的技术发明的应用领域也越来越多,可见数据挖掘相关技术领域的学科发现对其技术发明有着积极的影响。多学科、多领域的科学发现与技术的融合,必然促进数据挖掘或其他技术的创新与应用,推动相关知识理论的发展。

参考文献:

[1] 董坤,许海云,罗瑞,等. 科学与技术的关系分析研究综述[J]. 情报学报,2018,37(6):642 – 652.

[2] 王刚波, 官建成. 纳米科学与技术之间的联系:基于学术型发明人的分析[J]. 中国软科学, 2009(12): 71 – 79.

[3] KESSLER M M. Bibliographic coupling between scientific papers [J]. American documentation, 1963,14(1):10 – 25.

[4] BIKARD M. Made in academia: the effect of institutional origin on inventors' attention to science[J]. Organization science, 2018,29(5): 818 – 836.

[5] AHMADPOOR M, JONES B F. The dual frontier: patented inventions and prior scientific advance[J]. Science, 2017,357(6351): 583 – 587.

[6] HUANG M H, YANG H W, CHEN D Z. Increasing science and technology linkage in fuel cells: a cross citation analysis of papers and patents[J]. Journal of informetrics, 2015,9(2):237 – 249.

[7] DING C G, HUNG W C, LEE M C, et al. Exploring paper characteristics that facilitate the knowledge flow from science to technology[J]. Journal of informetrics, 2017,1(1):244 – 256.

[8] WANG X, ZHAO Y, LIU R, et al. Knowledge-transfer analysis based on co-citation clustering[J]. Scientometrics, 2013,97(3): 859 – 869.

[9] LIU G. Visualization of patents and papers in terahertz technology: a comparative study[J]. Scientometrics, 2013,94(3):1037 – 1056.

[10] GAO J P, TENG L, PANG J. Hybrid documents co-citation analysis: making sense of the interaction between science and technology in technology diffusion[J]. Scientometrics, 2012,93(2):459 – 471.

[11] QI Y, ZHU N, ZHAI Y, et al. The mutually beneficial relationship of patents and scientific literature: topic evolution in nanoscience[J]. Scientometrics, 2018,115(2): 893－911.

[12] RISCH J, KRESTEL R. What should I cite? cross-collection reference recommendation of patents and papers[C]//KAMPS J, TSAKONAS G, MANOLOPOULOS Y, et al. 21st international conference on theory and practice of digital libraries. Cham: Springer, 2017: 40－46.

[13] GOLNABI H. Carbon nanotube research developments in terms of published papers and patents, synthesis and production[J]. Scientia iranica, 2012,19(6):2012－2022.

[14] WONG C Y, GOH K L. The sustainability of functionality development of science and technology: papers and patents of emerging economies[J]. Journal of informetrics, 2012,6(1):55－65.

[15] PRATHAP G. Totalized input-output assessment of research productivity of nations using multi-dimensional input and output[J]. Scientometrics, 2018, 115(1): 577－583.

[16] 杜建,孙轶楠,李永洁,等. 从科学－技术交叉处识别创新前沿:方法与实证[J]. 情报理论与实践,2019,42(1):94－99.

[17] 覃佳慧,何耶奇,叶鹰. 科学论文和技术专利的引用时滞及循环周期研究——以富勒烯为例[J]. 情报理论与实践,2018,41(7):23－25.

[18] 徐红姣,曾文,张运良. 基于 Word2vec 的论文和专利主题关联演化分析方法研究[J]. 情报杂志,2018,37(12):36－42.

[19] 曾文,徐红姣,李颖,等. 基于 VSM 的科技期刊文献与专利文献的相似度计算方法研究[J]. 情报工程,2016,2(3):37－42.

[20] 董坤,吴红. 基于论文－专利整合的3D打印技术研究热点分析[J]. 情报杂志,2014,33(11):73－76,61.

[21] 骆云中,陈蔚杰,徐晓琳. 专利情报分析与利用[M]. 上海:华东理工大学出版社,2007:14.

[22] 魏来,高霏霏. 专利发明人与申请人之间的合作关系研究[J]. 情报学报,2016,35(5):463－471.

[23] 王刚波,官建成. 纳米科学与技术之间的联系:基于学术型发明人的分析[J]. 中国软科学,2009(12):71－79.

[24] 邱均平. 论“引文耦合”与“同被引”[J]. 图书馆,1987(3):13－19.

[25] 叶春霞,余翔,李卫. 企业间专利合作的多学科知识网络研究[J]. 情报杂志,2013(4):113－120.

[26] 王曰芬,王金树,关鹏. 主题－主题关联的学科知识网络构建与演化分析[J]. 情报科学,2018,36(9):9－15,102.

[27] 刘勇,杜一. 网络数据可视化与分析利器:Gephi 中文教程[M]. 北京:电子工业出版社,2017:182－187.

[28] 魏来,高希然. 大数据背景下高校数据馆员的角色定位[J]. 情报资料工作,2015(5):90－94.

作者贡献说明:

宁子晨:进行数据调研,提出研究思路,撰写论文;

魏来:对论文的研究思路与内容撰写提供整体指导。

Research on the Relationship Between Patent Documents and Academic Papers Based on Patent Subjects ——A Case Study of Data Mining

Ning Zichen Wei Lai

School of Information Science and Technology, Northeast Normal University, Changchun, 130117

Abstract: [**Purpose/significance**] The patent documents and academic papers show the new progress of technological innovation and scientific research respectively. Combining patent documents with academic papers through patent subjects, and then conducting technical subject evolution analysis, which has certain reference significance for further discovering understanding the relationship between patent technology and scientific research. [**Method/process**] Taking the academic inventors in the field of data mining as the link, the association method was proposed and the research framework was constructed from 3 perspectives of patent subject-keyword coupling, IPC coupling and IPC-keyword co-occurrence. Then this paper analyzed the evolution of multi-dimensional relationship among subjects, technologies and themes in different time periods, and explored the relationship between patent documents and academic papers. [**Result/conclusion**] The role of academic inventors in the innovation of data mining is more and more important. The technical themes of most subjects are similar, and some even show a high degree of unity. However, there are also a few technologies that are not directly related to the themes, and the degree of difference is also large. But whether technology is directly related to the theme matter or not, technology inventions and scientific research have achieved deeper mutual penetration in the field of data mining.

Keywords: patent subject patent documents academic paper association discovery